The Potential of Flexible UL/DL Slot Assignment in 5G Systems

Catania, Davide; Gatnau, Marta; Cattoni, Andrea Fabio; Frederiksen, Frank; Berardinelli, Gilberto; Mogensen, Preben

Published in:
Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th

DOI (link to publication from Publisher):
10.1109/VTCFall.2014.6966202

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
https://doi.org/10.1109/VTCFall.2014.6966202

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Users may download and print one copy of any publication from the public portal for the purpose of private study or research.

Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.
The Potential of Flexible UL/DL Slot Assignment in 5G Systems

Davide Catania¹, Marta Gatnau Sarret¹, Andrea F. Cattoni¹, Frank Frederiksen², Gilberto Berardinelli¹, Preben Mogensen¹,²
¹Department of Electronic Systems, Aalborg University, Denmark
²Nokia, Denmark
dac@es.aau.dk, mgs@es.aau.dk, afc@es.aau.dk, frank.frederiksen@nsn.com, gb@es.aau.dk, pm@es.aau.dk

Abstract—5th Generation (5G) small cells are expected to satisfy the increasing demand for wireless data traffic. In the presence of large scale dense and randomly deployed cells, autonomous and distributed configuration mechanisms are highly desirable. However, small cells typically serve a small number of users, such that sudden traffic imbalances between downlink (DL) and uplink (UL) are expected in the new 5G system. We exploit the flexibility of time-division duplex (TDD) to deal with such imbalances by adapting swiftly to instantaneously varying traffic needs. In this paper we propose a distributed algorithm to deal with these varying traffic requirements. We also exploit the availability of interference rejection capable receivers. Simulation results show that in the presence of the aforementioned features, we can approximately double the session throughput and halve the packet delay in a large number of cases.

I. INTRODUCTION

The demand for high speed wireless data services is always on the rise. An exponential increase in traffic growth is expected in the coming years [1], paving the way for a future 5G system. One way of handling this traffic growth is via cell densification, and further spectral reuse [2]. We therefore expect the proliferation of a new wireless local area 5G system. While the exact requirements of a future 5G system might still be unclear, some key performance indicators to take into account include service availability, peak throughput, and latency.

Some envisioned characteristics include interference rejection capable receivers, and a redesigned frame structure. A direct impact on minimizing latency is to reduce the frame duration significantly as presented in [3]. Another envisioned feature is the usage of Time Division Duplex (TDD) mode, with complete freedom of assigning each frame as uplink or downlink. As opposed to frequency division duplex (FDD), TDD can offer advantages in terms of cost, possibilities of exploiting unpaired bands and coping with unbalanced uplink/downlink (UL/DL) traffic. Such a feature is particularly useful in local area scenarios, where the amount of active uplink users is typically larger and their traffic fluctuations have a minimal impact on the instantaneous UL/DL traffic needs of a cell. Long Term Evolution (LTE) [4] and WiMAX [5] have already included the concept of a switching point to deal with the aforementioned needs. In WiMAX, the switching point within a frame can be set dynamically to assign the appropriate amount of UL and DL timeslots. In LTE, a set of configurations are defined, and each cell can select one of these configurations allowing a downlink assignment varying from 40 to 90 percent of the available transmission time intervals (TTI) [4]. The reconfiguration time is limited by a timer [6] where studies in [7] show that lowering down such reconfiguration time can provide benefits to the system in terms of throughput, especially at lower loads. In our envisioned 5G system, we remove any switching point restrictions, such that each slot can be arbitrarily set to either uplink or downlink [3].

The key contribution of this paper lies in presenting a simple distributed UL/DL slot selection scheme showing the potential gains in session throughput and reduced packet delays when having complete control and freedom in assigning each slot as either UL or DL, as opposed to a fixed slot strategy, in the context of the envisioned 5G system. We also show that interference rejection combining (IRC) capable receivers are an effective way to suppress interference variations introduced by the proposed flexible UL/DL slot allocation scheme. The paper is organized as follows. In Section II we give a short overview on our envisioned 5G concept. In Section III we then proceed to introduce our proposed UL/DL selection scheme and its associated parameters. System level simulation results are shown in Section IV, and Section V finally concludes the paper and states the future work.

II. ENVISIONED 5G CONCEPT

In this section we present the most relevant concepts of our envisioned 5G concept related to this study. In the first subsection we will describe our envisioned frame structure, its direct impact on latency and the flexibility it provides in assigning each slot arbitrarily as UL or DL. Thereafter we will proceed to describe the potential of IRC receivers in the
context of our problem. A general overview of the whole concept can be found in [8].

A. Frame Structure

One key characteristic feature of the envisioned 5G frame format is its short 0.25ms duration. The main goal of using such a short frame is to reduce latency. This simplified frame structure is shown in Figure 1.

```
<table>
<thead>
<tr>
<th>DL Control</th>
<th>UL Control</th>
<th>DMRS</th>
<th>DATA (UL/DL/MUTE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t = 0.25ms</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Fig. 1. 5G Frame Structure

The frame consists of a downlink and uplink control part, a demodulation reference symbol (DMRS) and a data slot part. The data slot can be set to UL, DL or MUTE, giving us full flexibility to assign each slot arbitrarily every 0.25ms. Further details related to the frame design and the switching costs incurred from having such a short frame structure can be found in [3].

The DL control part is used by the access point (AP) to signal grant messages. A grant is essentially a transmission opportunity indication along with associated transmission parameters. It indicates whether the corresponding transmission is in DL or UL, and related information such as the Modulation and Coding Scheme (MCS), rank indicator (number of transmission streams), and allocated channels to be used.

A complete DL transmission procedure operates as follows. Let us consider an AP that decides to schedule a DL transmission towards a particular user equipment (UE). A grant is sent in the DL control part. On the following frame, the DL transmission towards the UE occurs. The one frame delay gives sufficient time to the UE receiving the grant to decode and process such information.

In the case of an UL transmission the following occurs. A scheduling request (SR) is sent by a UE to the AP in the UL control part, indicating its desire to be scheduled in UL. The AP decodes this SR and decides to grant the UE an UL transmission. Such a grant is sent on the subsequent frame, and the associated UL transmission occurs in data part, one frame after the grant is sent.

The DMRS symbol is used to enhance IRC operation, by allowing the receiver to estimate the interference covariance matrix that will be present in the data part [9].

B. Reducing interference variation via IRC

Another relevant key feature of our envisioned 5G system lies in the availability of interference rejection via IRC. Multiple input multiple output (MIMO) antenna systems offer extra degrees of freedom in the spatial domain. The antennas can be exploited by using multiple transmission streams, or by using the antennas to suppress interferers. In a system with more than two antennas a balance between the two is also possible.

In a fully flexible TDD UL/DL system, where each cell can switch its transmission direction independently from the neighbouring cells at each frame, the signal to noise plus interference ratio (SINR) experienced by a node is expected to vary quite wildly making it problematic for link and rank adaptation as well as for interference coordination schemes to converge. While the focus of this paper is on the analysis of the potential gains of flexible UL/DL slot selection rather than on rank adaptation, in this section we show that IRC receivers can be an extremely effective antidote to mitigate such a problem. IRC relies on an estimate of the interference covariance matrix for projecting the significant interferers onto an orthogonal subspace with respect to the desired signals. The system design in [8] enables an instantaneous estimate of the interference covariance matrix at each frame via the DMRS symbol, thus enabling the possibility of rejecting the active interferers from neighbouring cells regardless of their transmission direction.

Two important metrics that can severely impact system performance are the experienced average SINR, and the SINR variance. Intuitively we would like to increase the average SINR, and minimize the SINR variance. By minimizing the SINR variance, we ensure that the interference conditions felt by a particular node are independent of a neighbouring cell’s varying UL/DL slot selection. We simulate a set of 10 randomly deployed dense networks positioned in a 5 by 2 grid fashion, where each network consists of a single AP and a single UE. The simulation is repeated for different rank transmission schemes, e.g. Rank 1 transmission scheme - all nodes use a transmission rank of 1. We then analyse the SINR variation experienced by the different nodes when using IRC. The DL and UL traffic share is set to be equal, such that at a particular frame, each cell can equally go for either UL, or DL based on which direction has the most data to transmit.

A concise representation showing the SINR variation behaviour in the different schemes is shown in figure 2. Here we show the average SINR versus the average SINR variance for all nodes. The average SINR variance is extracted by computing the SINR variance for each node, and taking the average of this variance across all nodes for each deployment. It is quite clear that at rank 1, IRC can be effective in mitigating interference, by providing a high average SINR and low average SINR variance. The price to pay here is the inability to fully utilize all the resources via higher order spatial multiplexing, as all the available degrees of freedom are used to reject interference. On increasing the rank, the average SINR decreases, since fewer antennas are used to reject interference. At rank 2 and 3 the SINR variance grows significantly, since a neighbouring cell containing an interferer we cannot reject might be active or inactive based on the neighbour’s decision to schedule it or not. In summary, figure 2 shows that if needed, at rank 1, IRC is able to mitigate interference variation effectively.
Our proposed algorithm works as follows. We introduce two main parameters which impact the direction of transmission. These are a buffer size threshold and a Head-of-Line (HOL) delay threshold, denoted as $T_{bufferSize}$ and $T_{HOLDelay}$ respectively. The main role of $T_{bufferSize}$ is to avoid the buffer from overflowing once the buffer size starts growing excessively, while the role of $T_{HOLDelay}$ lies in bounding the delay experienced by a packet.

The AP is considered to be the decision maker, since it is the entity sending the grants. Internally the AP has updated information related to the buffer size and the current HOL delay in DL, but it needs to be informed of such metrics in UL. We assume that such UL information is ideally embedded in the SR sent from the UE to the AP. This allows the AP to have updated information related to buffer status and HOL delays in both UL and DL, allowing it to take a sensible decision.

The algorithm initializes its current direction to downlink by default. It then waits until it is requested a direction decision, on the start of a frame. Once requested, it checks whether there is any data in the UL or DL direction. If no data is to be transmitted in any of the two directions, the frame is muted. If data is available in only one direction, the algorithm will schedule the direction having data. If data is available in both directions, the algorithm will inspect the buffer size and HOL delay of the direction not being utilized. If any of the buffer size or HOL delay exceed the predefined thresholds, it switches the direction, otherwise it keeps the current direction. This operation is illustrated in figure 3.

IV. PERFORMANCE EVALUATION

In this section we present the results of the presented algorithm. To conduct our analysis, we use a custom discrete event based system level simulator. The most relevant simulation parameters are shown in Table I. Our scenario consists of a 10x2 grid consisting of 20 apartments each having a size of 10m x 10m as shown in figure 4. The apartments are separated by walls and in each apartment we deploy an AP and a single UE terminal, both of which are randomly deployed. The UE’s affiliate to an AP according to a closed subscriber group policy, meaning that a UE connects to the AP located in the same apartment.

Based on the analysis carried out in Section II-B, in order to lower the interference variation, a rank 1 configuration is used. This allows us to limit the inherent convergence problems of link and rank adaptation algorithms, allowing us to solely focus on the benefits of a flexible UL/DL allocation.

The packets generated during the on period of the bursty traffic model, are eventually fragmented into 1500 byte packets, representing the very common Ethernet v2 maximum transmission unit (MTU). All links use the same on-off traffic
model specified in Table I. The instantaneous packet arrival will temporarily overload the network but the radio link control (RLC) layer will buffer the packets. On average the traffic load will represent 30% load in one case and 70% in the second case. We also use a simplified RLC model where we simply aggregate and send as many packets as we can during a frame transmission opportunity. If the transmission opportunity provided by the MAC is much larger than the data available in the RLC buffers we simply use data padding. The buffer size was set to be infinite. While this is ideal and unrealistic in practical systems, such an assumption allows us to keep our focus on the merits of the algorithm, neglecting any external behaviours occurring due to buffer overflows.

Given our traffic model, we have found out that the performance of the algorithm is insensitive to the $Th_{BufferSize}$ threshold, since the algorithm was in most cases performing a switch in direction based on the $Th_{HOLDelay}$ threshold. To explain such behaviour, let us consider an arriving burst of data in UL, and an active burst of data in DL currently being served. In principle, setting a $Th_{BufferSize}$ threshold close to the average packet size, will allow the arriving burst of data to be served immediately, but the switch in the other direction will also happen quickly, since there is a high probability that the current experienced buffering delay already exceeds the defined $Th_{HOLDelay}$. Eventually, the direction which was triggered because of $Th_{BufferSize}$, will also experience buffering delay, and will subsequently switch because of the $Th_{HOLDelay}$ threshold. Setting a $Th_{BufferSize}$ threshold well below the average packet size expected in the system, will eventually make the algorithm switch direction on every slot, for any defined $Th_{HOLDelay}$. Such behaviour occurs because the algorithm will switch direction continuously until the buffer size goes under the defined threshold. However, when that point is reached, the buffering delay would have grown above any reasonably defined $Th_{HOLDelay}$, and hence a fast switch will occur independently of the defined $Th_{HOLDelay}$. On the other hand setting a large $Th_{BufferSize}$ will make the system switch direction based on the $Th_{HOLDelay}$ threshold. This proves that the $Th_{BufferSize}$ threshold does not significantly affect the algorithm outcome. Given our traffic model, instantaneously high loads were expected and all schemes ended up giving a similar performance when sweeping through different $Th_{HOLDelay}$ values. Similar performance was observed because at high bursty loads, the delay of a packet will eventually grow due to buffering delay and exceed any reasonably defined $Th_{HOLDelay}$. At this point, the algorithm will then converge to switch on every slot independently of the chosen $Th_{HOLDelay}$ parameter. The $Th_{HOLDelay}$ parameter was hence chosen to be very short to switch as fast as possible to reduce the individual packet delay.

As a baseline result we consider two different schemes. Firstly we assume a simple fixed one UL to one DL slot allocation, referred to as a Fixed Slot 1:1 strategy hereinafter. We also consider a traffic based scheme which simply allocates the transmission to the direction instantaneously having more data to transmit. If no data in either direction is present the frame is muted for both schemes.

In Figures 5, and 6 we compare the Fixed Slot 1:1, as well as the traffic based scheme to our proposed algorithm in terms of session delay and session throughput, respectively, when the system is loaded at 30% of the system’s maximum capacity.
A session is defined as a burst of data occurring during the on period of the traffic generator. The session throughput represents the amount of bits received from that session divided by the time taken to receive all the fragmented packets in the session, starting from the first received fragmented packet since the session is created. The session delay measures the amount of time taken to receive the last fragmented packet of the session, starting from the first received fragmented packet since the session is created. Additional delays incurred due to packet processing are neglected and excluded from the session delay.

As expected and witnessed by Figures 5 and 6 we observe gains in both session throughput and delay from our algorithm compared to the Fixed Slot 1:1 scheme. This is because whenever a sudden burst in traffic in one direction exceeding half of the cell’s capacity occurs, a rigid fixed slot allocation scheme can only accommodate half of the cell’s capacity while our algorithm can exploit the full cell capacity if the other direction is inactive. Similar benefits from the traffic based scheme are also observed in the session throughput in figure 5. The traffic based scheme is efficient in terms of exploiting the available capacity in the channel as it tends to maximize the link efficiency by sending as much data as it can, hence it is also able to reach the maximum system capacity in terms of session throughput. The problem with the traffic based scheme occurs when both UL and DL directions have data to transmit, and one direction is more loaded than the other. The less loaded direction, i.e. the direction having smaller amounts of data in its buffer, can be momentarily starved. This increases the packet delay as shown in figure 6 and also lowers the session throughput for that link as shown in the lower percentiles of the CDF in figure 5. The defined $T_{HOLDelay}$ in our algorithm ensures a degree of fairness and helps us bound the delay without letting it grow excessively unlike the traffic based scheme.

Tables II and III show some numerical results at 30% and 70% average load for the session delay, session throughput and final throughput. The final throughput is defined as the total amount of bits received over the total simulation time.

Ideally if the bursty nature of traffic allowed us to always instantaneously have only one direction active we would expect to double the session throughput and halve the packet delay, when compared to a Fixed Slot 1:1 scheme. However, since this is not always the case the observed gains from Tables II and III are slightly reduced to 49% for the delay and to 41% for the session throughput at 30% load for half of the cases. On increasing the load to 70%, and therefore increasing the probability of having two active links at a time, the gains of our proposed algorithm reduce to 41% in terms of delay for half of the cases.

In a multi-cell scenario there are cases where we can increase the gains even more than 50%. Even though IRC is effective at suppressing interference, there will be some nodes who are interfered by an amount of nodes that exceed the number of degrees of freedom dedicated to interference suppression. Since our algorithm services packet bursts more quickly, it can statistically reduce the interference levels over time, hence improving the conditions for these interfered nodes, allowing them to increase their session throughput significantly. This can be observed in the 5th percentile of the session throughput at 30% load.

Finally, having full UL/DL flexibility, allows us to also reach a higher peak final throughput when loading each cell with more than 50% load. This can be observed from Table III where the proposed algorithm gets an improvement of 8% in final throughput at 70% load compared to the Fixed Slot 1:1 scheme. This happens because the provided flexibility, unlike a rigid Fixed Slot 1:1 scheme, does not lock the UL and DL capacity to half the cell’s capacity.

V. CONCLUSIONS & FUTURE WORK

In this paper we have presented a simple and effective algorithm that exploits flexible UL/DL slot allocation in our
TABLE II
RESULTS FOR DIFFERENT LOADS IN PERCENTILES

<table>
<thead>
<tr>
<th></th>
<th>Fixed Slot</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>50%</td>
</tr>
<tr>
<td>30% Load</td>
<td>3.5</td>
<td>35</td>
</tr>
<tr>
<td>70% Load</td>
<td>6</td>
<td>70</td>
</tr>
<tr>
<td>Session Throughput (Mbps)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>50%</td>
<td>95%</td>
</tr>
<tr>
<td>30% Load</td>
<td>57</td>
<td>428</td>
</tr>
<tr>
<td>70% Load</td>
<td>13</td>
<td>219</td>
</tr>
<tr>
<td>Final Throughput (Mbps)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>50%</td>
<td>95%</td>
</tr>
<tr>
<td>30% Load</td>
<td>59</td>
<td>146</td>
</tr>
<tr>
<td>70% Load</td>
<td>179</td>
<td>305</td>
</tr>
</tbody>
</table>

TABLE III
ALGORITHM GAINS OVER FIXED SLOT STRATEGY

<table>
<thead>
<tr>
<th></th>
<th>Gains in %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
</tr>
<tr>
<td>30% Load</td>
<td>42.9</td>
</tr>
<tr>
<td>70% Load</td>
<td>41.7</td>
</tr>
<tr>
<td>Session Throughput (Mbps)</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>50%</td>
</tr>
<tr>
<td>30% Load</td>
<td>63.0</td>
</tr>
<tr>
<td>70% Load</td>
<td>43.5</td>
</tr>
<tr>
<td>Final Throughput (Mbps)</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>50%</td>
</tr>
<tr>
<td>30% Load</td>
<td>-3.5</td>
</tr>
<tr>
<td>70% Load</td>
<td>-3.5</td>
</tr>
</tbody>
</table>

envisioned 5G system. The algorithm is capable to automatically track traffic variations, and adapt accordingly to the instantaneous needs of the cell. The flexibility of the envisioned frame structure gives us the possibility to adapt to sudden traffic imbalances between UL and DL and allows us to exploit the full capacity of our system rather than being constrained by a rigid fixed allocation scheme. Compared to a fixed slot scheme with one slot in uplink and the other in downlink we are capable of almost doubling the experienced session throughput and reducing the experienced delay by half in approximately 50% of the cases. We are also able to reduce the delay compared to a traffic based scheme. This is achieved by introducing a head of line delay parameter threshold. We also analyse and confirm that interference rejection via IRC is an effective tool in increasing robustness from the source of interference and can help us in stabilizing the experienced interference, hence counteracting and minimizing previous challenges experienced with flexible UL/DL slot allocation.

Our future work will focus on exploiting spatial multiplexing gains while limiting SINR variations via appropriate link and rank adaptation algorithms. The impact of hybrid automatic repeat request (HARQ) on reducing packet losses will also be studied and its impact on the delay will be investigated. The behaviour of such an algorithm in the presence of multiple users requires further investigation and the impact of reduced signalling and the compression of feedback reports needs to be quantified. Finally, the behaviour of the algorithm with different parametrizations, in the presence of higher layer protocols such as the Transmission Control Protocol (TCP), also needs to be assessed, as such protocols provide the possibility of testing the algorithm in instantaneously high asymmetric conditions due to the presence of TCP acknowledgements.

REFERENCES